LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-584853

Level-2 Milestone 4470: Early Users on Unclassified Sequoia Hardware

Milestone report for NNSA HQ

Prepared by David Fox
September 14, 2012

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Contents

Introduction

This report documents the delivery and installation of Sequoia and that early users have successfully run codes on the Sequoia machine as part of the requirements for ASC L2 Milestone 4470: Early Users on the Unclassified Sequoia Hardware, due September 30, 2012. The full text of the milestone is included in Attachment 1. The description of the milestone is:

> *The Sequoia final system is planned to provide peak compute power of over 20 petaFLOP/s of computing cycles to the weapons program. The Sequoia system will be delivered in phases beginning in the first quarter of FY12, with final rack deliveries early in the third quarter of FY12. The goal of this milestone is to have early users successfully running their codes on some portion of the Sequoia system on the unclassified network.*

The milestone was completed June 14 2012, when the ALE3D code first successfully ran on the Sequoia machine. The following sections describe the Sequoia system architecture, current status, and installation time line.

A project plan is included as Attachment 2. Attachment 3 is a letter certifying that the user successfully ran on the Sequoia system.

Sequoia System Architecture

The Sequoia machine is an IBM BlueGene/Q system composed of 96 compute racks. Each compute rack has 1024 compute nodes and 8 I/O nodes. Both compute nodes and I/O nodes have 16 Power A2 processor cores and 16 GB of memory. Each compute core is capable of running 4 MPI tasks as hardware threads on the core. The system has a total of 1,572,864 compute cores supporting up to 6,291,456 hardware compute threads, 768 I/O nodes, 1.5PB total memory, and 20.1 PF peak performance. Figure 1 shows the major components of the Sequoia system.

End-users access to the system is via various Login Nodes. Sequoia has 2 Login Large Application Compile nodes (LN-LAC) each with 12 Power 7 cores and 64GB RAM, and 16 Login/Application Development Nodes (LN-AD) each with 16 Power A2 cores and 16GB RAM. The Login Nodes are used to compile, run, and monitor compute jobs and access file system resources.

Each of the 768 I/O nodes provides I/O services for 128 compute nodes. The I/O nodes mount the NFS and Lustre file systems. The Lustre file system deployed along with Sequoia is 50 PB in size and is targeted to delivery at least 500 GB/s of peak performance.

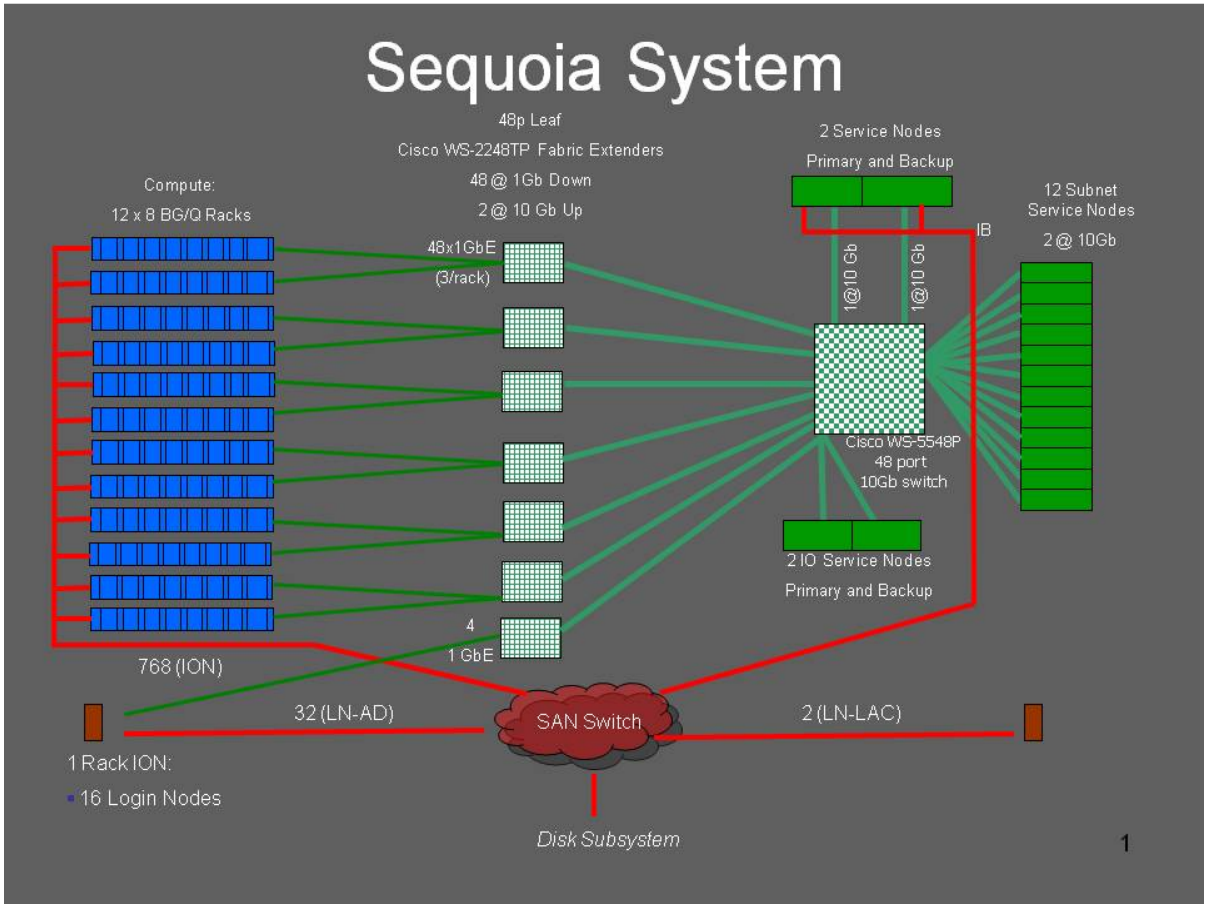

Figure 1: Sequoia Architecture

Installation Time Line

The preparation of the Terascale Simulation Facility (TSF) for the Sequoia machine started with the retirement of the Purple system in November 2010. The decommissioning and removal of the Purple machine and all the under floor support was completed in July 2011. This allowed for the construction of the chilled water system, the floor supports and power infrastructure to support the Sequoia machine. The construction of the support systems for Sequoia was a challenge mainly because Sequoia was the first water cooled system installed in the TSF. For example, the final commissioning of the water system was delayed till February 3rd 2012 due to contaminants in the pipes, even though the first four racks of Sequoia were delivered January 3rd 2012.

The delivery of the 96 Sequoia racks happened over the first four months of 2012 with the last racks delivered April 16th. In the period between January and April IBM and LLNL people worked closely to coordinate the installation of racks and allow for time to run applications on the system. The running of application codes on the hardware as soon as possible was key to finding early life hardware failures and finding issues with the system software. Time was also allocated to early users to port and test application codes on the system.

Below are some major milestones tracked as part of the Sequoia install:

- March 21 2012 the first 24 racks complete a Linpack run
- April 3 2012 the second 24 racks complete a Linpack run
- April 21 2012 the first 48 racks complete a Linpack run
- May 2 2012 the Lammps benchmark runs on 3.1 million MPI tasks
- May 14 2012 the first 72 racks complete a Linpack run
- June 8 2012 the full system Linpack run completes
- June 14 2012 Sequoia is announced as the #1 system on the Top500 Super Computer list

Additional Applications run on Sequoia

In addition to the code used to certify this milestone (ALE3D), a number of additional application codes have been ported and run on the system. Below is a list of some of the major applications that have been run (Table 1).

Code/Project	Total Nodes	MPI Tasks/Node	Compute Threads/Task	Total Compute Threads
Lattice Gauge Theory	96K	1	64	6.2M
Cardioid – Simulation of the electrophysiology of the human heart (Gordon Bell Award Finalist)	48K	1	64	3.1M
HACC – Extreme Scale Computational Cosmology (Gordon Bell Award Finalist)	48K	8	8	3.1M
Qbox – a first-principles molecular dynamics code	24K	1	32	768K
ALE3D - Arbitrary Lagrangian-Eulerian 3D	8K	2	1	16K
AutoDock Vina – a molecular docking code	8K	8	1	64K

Table 2: Application Codes on Sequoia

Attachment 1: Milestone Definition Text

<table>
<tr><th colspan="6">Milestone (ID#4470): Early Users on Unclassified Sequoia Hardware</th></tr>
<tr><td colspan="2">Level: 2</td><td colspan="2">Fiscal Year: FY12</td><td colspan="2">DOE Area/Campaign: ASC</td></tr>
<tr><td colspan="6">Completion Date: 9/31/12</td></tr>
<tr><td colspan="6">ASC nWBS Subprogram: CSSE/FOUS</td></tr>
<tr><td colspan="6">Participating Sites: LLNL</td></tr>
<tr><td colspan="6">Participating Programs/Campaigns: ASC</td></tr>
<tr><td colspan="6">Description: The Sequoia final system is planned to provide peak compute power of over 20 petaFLOP/s of computing cycles to the weapons program. The Sequoia system will be delivered in phases beginning in the first quarter of FY12, with final rack deliveries early in the third quarter of FY12. The goal of this milestone is to have early users successfully running their codes on some portion of the Sequoia system on the unclassified network.</td></tr>
<tr><td colspan="6">Completion Criteria: Racks have been assembled in the TSF and have run at least one user’s ported code. A user will write a memo certifying that their code has run on Sequoia.</td></tr>
<tr><td colspan="6">Customer: ASC</td></tr>
<tr><td colspan="6">Milestone Certification Method:
A report will be written as a record of milestone completion. The memo from the user certifying that his/her code has run will also be submitted.</td></tr>
<tr><td colspan="6">Supporting Resources: ASC program people, the Sequoia system, IBM</td></tr>
<tr><td colspan="6">Supporting Milestones:
<table>
<tr><th>Program</th><th>Title</th><th>Due Date</th></tr>
<tr><td>N/A</td><td></td><td></td></tr>
</table></td></tr>
<tr><td colspan="6">Codes/Simulation Tools Employed: Sequoia software environment</td></tr>
<tr><td colspan="6">Contribution to the ASC Program: Advanced architecture resource for UQ, 20 petaFLOP/s peak capability</td></tr>
<tr><td colspan="6">Contribution to Stockpile Stewardship: Acceleration of ability to do UQ.</td></tr>
<tr><td rowspan="2">No.</td><td rowspan="2">Risk Description</td><td colspan="4">Risk Assessment (low, medium, high)</td></tr>
<tr><td>Consequence</td><td>Likelihood</td><td colspan="2">Exposure</td></tr>
<tr><td>1.</td><td>Hardware issues</td><td>Moderate</td><td>Low</td><td colspan="2">Low</td></tr>
</table>

Attachment 2: Project Plan

Task Name	Duration	Start	Finish	Predecessors	% Complete
Sequoia Site Prep	**290 days**	**Mon 1/3/11**	**Fri 2/10/12**		**100%**
Floor/Tray	**288 days**	**Mon 1/3/11**	**Wed 2/8/12**		**100%**
Cooling	**290 days**	**Mon 1/3/11**	**Fri 2/10/12**		**100%**
Power	**137 days**	**Mon 8/1/11**	**Wed 2/8/12**		**100%**
IB SAN Prep	**10 days**	**Mon 12/19/11**	**Fri 12/30/11**		**100%**
Sequoia Installation and Integration	**362 days?**	**Wed 2/9/11**	**Thu 6/28/12**		**100%**
1/2 rack	**36 days**	**Mon 12/19/11**	**Mon 2/6/12**		**100%**
IB SAN at Rochester	**101 days**	**Mon 5/16/11**	**Mon 10/3/11**		**100%**
Initial 4 Rack System at LLNL	**29 days**	**Tue 1/3/12**	**Fri 2/10/12**		**100%**
Phase 3A Scaling System 8.3.24.1 (16 racks and mini-SWL at Rochester)	4 days	Mon 3/12/12	Thu 3/15/12	43	100%
Initial 24 Rack System at LLNL	**291 days**	**Wed 2/9/11**	**Wed 3/21/12**		**100%**
Delivery - racks 5-8	1 day	Mon 1/30/12	Mon 1/30/12		100%
Install racks 5-8	10 days	Mon 1/30/12	Fri 2/10/12	53FS-1 day	100%
Alpha access to 8 rack system	0 days	Fri 2/10/12	Fri 2/10/12	54	100%
Deliver racks 9-12	1 day	Tue 1/3/12	Tue 1/3/12		100%
Install racks 9-12	10 days	Tue 1/3/12	Mon 1/16/12	56FS-1 day	100%
Rework 50 Bulk Power Enclosures	3 days	Wed 2/9/11	Fri 2/11/11		100%
Deliver racks 13-21	1 day	Mon 2/13/12	Mon 2/13/12		100%
Install racks 13-21	10 days	Mon 2/13/12	Fri 2/24/12	59FS-1 day	100%
Deliver racks 22-26	1 day	Wed 2/15/12	Wed 2/15/12	59FS+1 day	100%
Install racks 22-26	10 days	Wed 2/15/12	Tue 2/28/12	61FS-1 day	100%
Integrate 24 rack system	1 wk	Wed 2/29/12	Tue 3/6/12	62	100%
24 Rack Linpack (Phase 1 System 8.3.24.2)	11 days	Wed 3/7/12	Wed 3/21/12	63	100%
Release for LLNL testing and users	0 days	Wed 3/21/12	Wed 3/21/12	64	100%
Initial 48 Rack System at LLNL	**43 days?**	**Wed 2/22/12**	**Fri 4/20/12**		**100%**
Deliver - racks 27-29	1 day	Wed 2/22/12	Wed 2/22/12		100%
Install racks 27-29	10 days	Wed 2/22/12	Tue 3/6/12	67FS-1 day	100%
Delivery racks 30-33	1 day	Fri 2/24/12	Fri 2/24/12	67FS+1 day	100%
Install racks 30-33	10 days	Fri 2/24/12	Thu 3/8/12	69FS-1 day	100%
Deliver racks 34-38	1 day	Thu 3/1/12	Thu 3/1/12	69FS+3 days	100%
Install racks 34-38	10 days	Thu 3/1/12	Wed 3/14/12	71FS-1 day	100%
Deliver racks 39-43	1 day?	Thu 3/8/12	Thu 3/8/12	71FS+4 days	100%
Install racks 39-43	10 days	Thu 3/8/12	Wed 3/21/12	73FS-1 day	100%
Deliver racks 44-48	1 day	Mon 3/12/12	Mon 3/12/12	73FS+1 day	100%
Install racks 44-48	10 days	Mon 3/12/12	Fri 3/23/12	75FS-1 day	100%
Integrate 2nd 24rack system	1 wk	Mon 3/26/12	Fri 3/30/12	76	100%
24 Rack Linpack (Phase 2 System 8.3.24.3)	3 days	Mon 4/2/12	Wed 4/4/12	77	100%
Integrate 48 rack system	1 wk	Thu 4/5/12	Wed 4/11/12	78	100%
48 Rack Linpack	7 days	Thu 4/12/12	Fri 4/20/12	79,221	100%
48 Rack SWL (Sequoia 48 System Acceptance 8.3.24.4)	9 days	Mon 6/18/12	Thu 6/28/12	80FS+2 mons	100%
Sequoia 2nd 48 Racks	**32 days**	**Thu 3/15/12**	**Fri 4/27/12**		**100%**
Deliver - racks 49-52	1 day	Thu 3/15/12	Thu 3/15/12		100%
Install racks 49-52	10 days	Thu 3/15/12	Wed 3/28/12	83FS-1 day	100%
Delivery racks 53-57	1 day	Fri 3/16/12	Fri 3/16/12	83	100%
Install racks 53-57	10 days	Fri 3/16/12	Thu 3/29/12	85FS-1 day	100%
Deliver racks 58-62	1 day	Wed 3/21/12	Wed 3/21/12	85FS+2 days	100%
Install racks 58-62	10 days	Wed 3/21/12	Tue 4/3/12	87FS-1 day	100%
Deliver racks 63-66	1 day	Fri 3/23/12	Fri 3/23/12	87FS+1 day	100%
Install racks 63-66	10 days	Fri 3/23/12	Thu 4/5/12	89FS-1 day	100%
Deliver racks 67-71	1 day	Thu 3/29/12	Thu 3/29/12	89FS+3 days	100%
Install racks 67-71	10 days	Thu 3/29/12	Wed 4/11/12	91FS-1 day	100%
Deliver racks 72-75	1 day	Fri 3/30/12	Fri 3/30/12	91	100%

Install racks 72-75	10 days	Fri 3/30/12	Thu 4/12/12	93FS-1 day	100%
Deliver racks 76-80	1 day	Mon 4/2/12	Mon 4/2/12	93	100%
Install racks 76-80	10 days	Mon 4/2/12	Fri 4/13/12	95FS-1 day	100%
Deliver racks 81-84	1 day	Wed 4/4/12	Wed 4/4/12	95FS+1 day	100%
Install racks 81-84	10 days	Wed 4/4/12	Tue 4/17/12	97FS-1 day	100%
Deliver racks 85-89	1 day	Tue 4/10/12	Tue 4/10/12	97FS+3 days	100%
Install racks 85-89	10 days	Tue 4/10/12	Mon 4/23/12	99FS-1 day	100%
Deliver racks 90-94	1 day	Thu 4/12/12	Thu 4/12/12	99FS+1 day	100%
Install racks 90-94	10 days	Thu 4/12/12	Wed 4/25/12	101FS-1 day	100%
Deliver racks 95-96	1 day	Mon 4/16/12	Mon 4/16/12	101FS+1 day	100%
Install racks 95-96	10 days	Mon 4/16/12	Fri 4/27/12	103FS-1 day	100%
Sequoia Integration and Linpack	**41 days**	**Fri 4/13/12**	**Fri 6/8/12**		**100%**
Integrate 72 rack system	1 wk	Fri 4/13/12	Thu 4/19/12	94	100%
72 rack linpack	17 days	Fri 4/20/12	Mon 5/14/12	106	100%
Integrate 96 rack system	1 wk	Tue 5/15/12	Mon 5/21/12	80,107	100%
96 rack Linpack #1 on Top 500	14 days	Tue 5/22/12	Fri 6/8/12	108	100%

Attachment 3: Certification Letter

ADVANCED SIMULATION & COMPUTING

TO: LLNL ASC Office

FROM: Fady M. Najjar

SUBJECT: **Completion of LLNL ASC Level 2 Milestone 4470**

As an end user of the Sequoia system, I certify that the Arbitrary Lagrangian-Eulerian 3D (ALE3D) code ran successfully on the Sequoia system to satisfy in part the requirements of the LLNL ASC Level 2 Milestone 4470, "Early Users on Unclassified Sequoia Hardware".

During the install and integration phase of the Sequoia machine, the ALE3D code was ported to the BlueGene/Q platform and run on the unclassified Sequoia machine. The ALE3D code was used in a scaling study up to 8K nodes with 16K total MPI tasks. In addition as part of the scaling study, the Livermore Unstructured Lagrangian Explicit Shock Hydrodynamics (LULESH) mini-app (which is derived from ALE3D) code was run on Sequoia on 48K nodes with a total of 3.1M MPI tasks.

The details of the efforts to install the Sequoia machine are documented in "Level 2 Milestone 4470: Early Users on the Unclassified Sequoia Hardware" report.

Date signed:
09/14/2012

Fady M. Najjar, Ph.D.

Code Physicist

Lawrence Livermore National Laboratory